STUDIA PHONOLOGICA XXIII (1989)

Speaker-Independent Consonant Recognition by Integrating Discriminant Analysis and HMM

Tatsuya KAWAHARA, Shuji DOSHITA and Shigeyoshi KITAZAWA

ABSTRACT

In this paper, we propose a new consonant recognition method which integrates two stochastic method : discriminant analysis and HMM (Hidden Markov Models). Discriminant Analysis is effective to analyze local patterns around the reference-point of a consonant such as a burst point. This method, however, is based on the assumption that the reference-point is detected precisely. HMM is able to extract the global dynamic features of a consonant from the preceding vowel to the following vowel and needs no explicit segmentation of speech. But it is hard to discriminate between similar consonants with HMM due to the quantization of input pattern vectors. Our new method constructs HMM with discriminant analysis front-end and recognizes consonants by combining the score obtained by discriminant analysis and the score by HMM. In recognition experiments of all the Japanese consonants in mono-syllables, this integrated method achieved the recognition rate of 92.1 %, which is higher by 5～15 % than the case using either of two methods alone.

1. INTRODUCTION

In order to realize large-vocabulary speaker-independent automatic speech recognition, phoneme-based recognition is desirable. Therefore we are studying recognition of consonants which is, due to their dynamic features, more difficult than that of vowels. There exist many approaches to consonant recognition. Among them statistical or probabilistic method is advantageous because it can avoid extracting explicit distinctive features and realizes a simple and flexible interface with the natural language processing unit.

Discriminant analysis, which is one of the multi-variate statistical analyses, is suitable to discriminate local patterns around the reference-point of a consonant

Tatsuya KAWAHARA (河原達也) : Doctoral course student, Department of Information Science, Faculty of Engineering, Kyoto University
Shuji DOSHITA (堂下修司) : Professor, Department of Information Science, Faculty of Engineering, Kyoto University
Shigeyoshi KITAZAWA (北澤茂良) : Associate professor, Department of Computer Science, Faculty of Engineering, Shizuoka University

such as a burst point or a starting point of friction. This method assumes the exact detection of the reference-point of consonants and such a precise segmentation of speech is extremely difficult.

On the other hand, HMM is able to extract global dynamic features of a consonant from the preceding vowel to the following vowel and it does not need precise segmentation. With conventional HMM, however, it is hard to discriminate similar consonants because quantizing input pattern vectors causes loss of discriminant information and the learning algorithm does not necessarily separate all the classes.

As we reviewed above, discriminant analysis and HMM are different in feature extraction and expected to be compatible. We, therefore, propose a new recognition method which integrates these two. It extracts the global features of a consonant with HMM and analyze local and detailed features with discriminant analysis. The final result is obtained by combining the scores calculated by these two methods. In this paper, we discuss on the basic concept of this recognition method and its implementation and the experimental results.

2. RECOGNITION WITH DISCRIMINANT ANALYSIS

In order to extract features independent of speakers and environments, we adopt one of the multi-variate statistical analyses, discriminant analysis[1]. Suppose population of class $\mathbf{i}$ is normally distributed with mean $\mathbf{u}_i$ and covariance $\sum_i$, and suppose further covariance matrices $\sum_i (i=1, \cdots, n)$ are equal to $\sum$. The probability density that a given pattern vector $\mathbf{x}$ belongs to a class $\mathbf{i}$ is as follows :

$$p(\mathbf{x}/\mathbf{i}) = C \cdot exp\{-D_i^2(\mathbf{x})/2\}$$

$$D_i^2(\mathbf{x}) = (\mathbf{x}-\mathbf{u}_i)^t \cdot \sum^{-1} \cdot (\mathbf{x}-\mathbf{u}_i)$$

$$C = \frac{1}{(2\pi)^{d/2} \cdot |\sum|^{1/2}}$$

where d is the dimension of $\mathbf{x}$ and $D_i^2(\mathbf{x})$ is called Mahalanobis distance.

The mean of each class and the covariance are estimated with training samples. Here statistical variable selection is performed so as to separate all the classes and reduce the dimension of the input vector. A given sample $\mathbf{x}$ will be classified into such a class $\mathbf{i}$ that the probability density $\mathbf{p}(\mathbf{x}/\mathbf{i})$ is the largest.

In consonant recognition, an input vector consists of some series of short-term spectra. In the training phase, such vectors are obtained by analyzing some consecutive frames around the reference-point specified by human observation. In the recognition phase, however, it is not practical to specify such points manually and automatic detection of them is extremely difficult.

To avoid explicit segmentation of speech, we apply the well-trained phoneme

classifier to every frame of speech. As a result, a sequence of phoneme-like symbols with their scores is gotten. The final result is obtained by processing this sequence, for example, choosing the symbol with the highest score. This process is illustrated in Fig. 1.

In this processing, a difficulty arises because, even if the correct discrimination is done at the very reference-point, the classifier often gives rather high scores to incorrect consonant symbols at frames distant from that point. Since the score calculated by discriminant analysis is based on the local features, judging with this score alone might cause a lot of insertion errors or incorrect choices from the sequence. In order to eliminate such errors, it is necessary to grasp the global features of consonants.

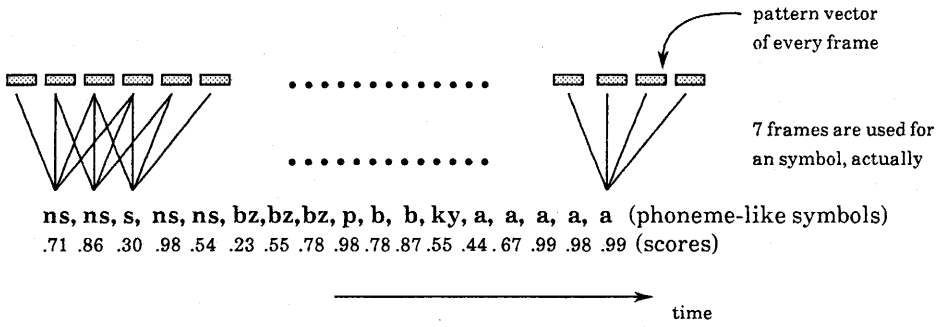

Fig. 1 Extracting a sequence of phoneme-like symbols and their scores by the classifier's scanning of speech

3. Recognition with HMM

In order to extract the global dynamic features of consonants, we introduce HMM. Each consonant is formulated by a left-to-right Markov model of several states and real speech is modeled as a sequence of symbols which a Markov model output in transiting its state at every frame. The parameters of models such as state transition probabilities and symbol output probabilities are estimated with training samples of sequences. A given sample $\mathrm{O}=(O_1, \cdots O_T)$ is classified into such a class M that the probability $p(\mathrm{O}/\mathrm{M})$ outputting O is the largest. The probability $p(\mathrm{O}/\mathrm{M})$ is calculated by the following:

a_{ij}: a probability of making a transition from state i to state j

$b_i(k)$: a probability of outputting symbol k at state i

$\alpha_1(1)=b_1(O_1), \quad \alpha_1(i)=0 \qquad 2\leq i\leq N, \quad N$: number of states

$$\alpha_t(i)=[\sum_{j=1}^{N}\alpha_{t-1}(j)\cdot a_{ji}]\cdot b_i(O_t) \qquad 2\leq t\leq T, \quad T \text{ : length of O}$$

$$p(\mathrm{O}/\mathrm{M})=\sum_{i=1}^{N}\alpha_T(i)$$

In the conventional HMM, an input consists of discrete symbols or codes, which are determined by VQ (vector quantization). VQ reduces the amount of computation and storage compared with treating pattern vectors directly, but quantization error is inevitable. This error is nothing but the loss of necessary information contained in the pattern vectors and lowers recognition performance especially in speaker-independent recognition. Using a large code-book may reduce the quantization error, but increases the number of parameters of HMM in square order, consequently makes it difficult to estimate them accurately.

In addition, the training algorithm commonly used lacks the concept of maximizing the distance between the classes although it can constructs an optimal model for each class. Some algorithms to conquer this defect have been proposed [3] but they are not adequate for the reason that convergence and the positiveness of probabilities are not guaranteed and huge training time is needed.

HMM, therefore, lacks the ability to discriminate acoustically similar consonants although it is effective to grasp global features.

4. INTEGRATING DISCRIMINANT ANALYSIS AND HMM

4. 1 Recognition by integrating discriminant analysis and HMM

Now we review the two methods described in the previous sections. Comparison is done on Table 1. While discriminant analysis extracts the local features using pattern vectors, HMM grasps the global dynamic features using symbol sequences. While discriminant analysis treats input variables as a vector, HMM regards an input sequence as a Markov chain.

Table 1 : comparison between discriminant analysis and HMM

	discriminant analysis	HMM
input	spectrum	symbol or code
	vectors	Markov sequences
feature extraction	local	global
	detailed	general
	combinatorial	dynamic

As we see, discriminant analysis and HMM are different in feature extraction and seem to be compatible. We, therefore, propose a new recognition method which integrates these two. The procedure of recognition is as follows. Fig. 2 is a flowchart of our method.

1. At every frame of speech, pattern vector is obtained by acoustic analysis.
2. The phonome classifier based on discriminant analysis is applied to every frame to get a sequence of phoneme-like symbols with their scores.

3. A score is calculated by checking the phonome-like symbol sequence with each HMM which represents a consonant.
4. Another score is obtained by summing the scores by discriminant analysis in the sequence for each consonant.
5. For each consonant, the product of the score of 3 and that of 4 is calculated.
6. The given sample is classified into a consonant which gets the largest product score.

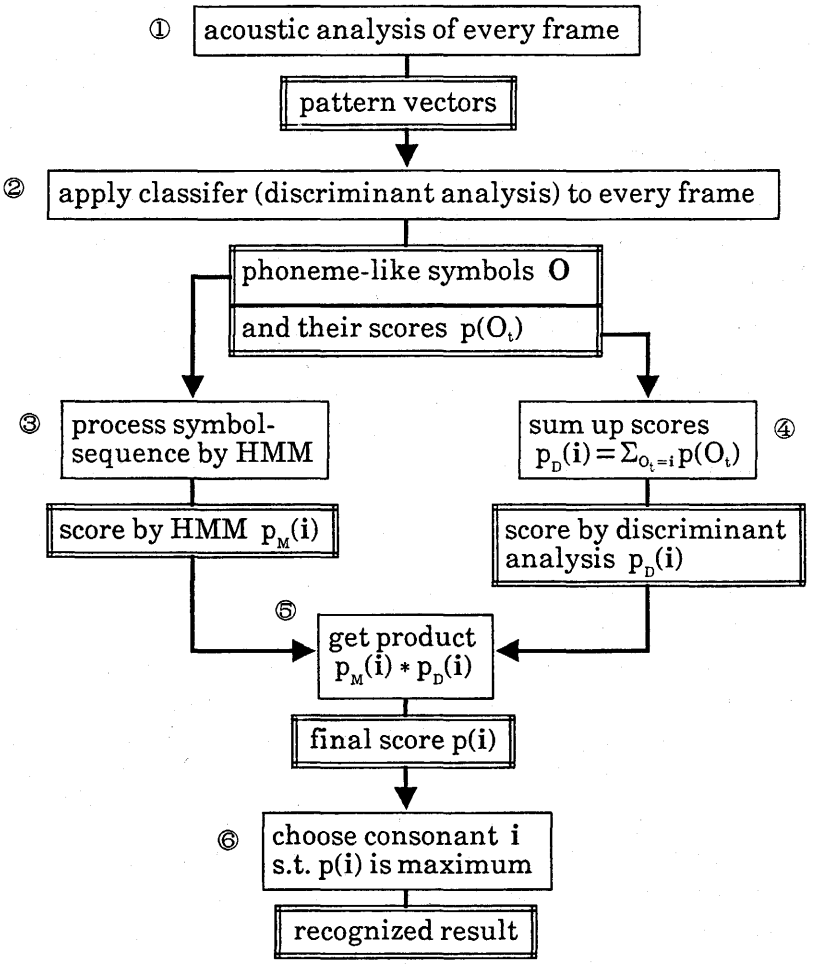

Fig. 2 Flowchart of recognition by integrating discriminant analysis and HMM

4. 2 The meaning of integrating discriminant analysis and HMM

In this section, we discuss the meaning of integrating discriminant analysis and HMM, and the meaning of multiplying the scores of the two methods from following two points of view.

(1) Improvement of discriminant analysis by HMM

As discussed in Section 2, the defect of recognition with discriminant analysis is that the classifier causes unexpected errors at frames distant from the reference-point. It is necessary to observe wider range of speech than the classifier's scope and suppress such errors using global context. To this purpose, the sequence is checked with HMM's which describes the features from the preceding vowel to the following vowel.

This post-processing gives not only discriminant information but also locative information. A phoneme symbol which is inadequate from the context of the sequence will be ignored even if it gets a high score. For example, if nasal-murmur symbols appear beforehand, the probabilities of consonants except nasals will be lowered. Our method is, therefore, a kind of improvement of discriminant analysis by HMM.

From this point of view, the meaning of the product of two scores is the combination of the score on locative information by HMM and the score on discriminant information by discriminant analysis.

$$p(\mathrm{i}\,;\ \textit{reference-point}) = p_M(\textit{reference-point}) * p_D(\mathrm{i}/\textit{reference-point})$$

where , $p_M(\textit{reference-point})$ is a probability, obtained by HMM, that there exists a reference-point of consonant i, and $p_D(\mathrm{i}/\textit{reference}\text{-point})$ is a probability, obtained by discriminant analysis, that a given pattern belongs to a consonant i supposing that it is a pattern of the reference-point, and $p(\mathrm{i}\,;\ \textit{reference}\text{-point})$ is a joint probability of them.

(2) Improvement of HMM by discriminant analysis

As discussed in Section 3, one of the defect of HMM is due to quantization error of VQ. Here we use phoneme-like symbols instead of VQ codes as front-end. Indeed the set of phonem-like symbols is also regarded as quantization of spectra, but these symbols have discriminant information concerning the reference-points. By utilizing this information, namely the scores by discriminant analysis, we realizes more precise discrimination than by treating only symbol sequences. Furthermore, statistical variable selection in discriminant analysis maximizes the distance between symbols, consequently improves the separatability of the classes to be recognized.

From this point of view, the meaning of the product of two scores is the combination of information on the global features by HMM and information on the local features by discriminant analysis.

$$p(\textit{global \& local features} / \mathrm{i}) = p_M(\textit{global features} / \mathrm{i}) * p_D(\textit{local features} / \mathrm{i})$$

where, $p_M(\textit{global features} / \mathrm{i})$ is a probability of consonant i after analyzing the global features with HMM, $p_D(\textit{local features} / \mathrm{i})$ is a probability of consonant i after analyzing the local features with discriminant analysis, and $p(\textit{global \& local features} / \mathrm{i})$ is obtained as a joint probability of them.

Another approach which does not use VQ is continuous-parameter HMM[2]. Our method has merit that training using discriminant analysis is easy and discriminant information around the reference-point, where the distinctive feature concentrates, can directly affects the whole recognition.

5. Consideration for Implementation

5. 1 Phoneme classifier based on discriminant analysis

The phoneme classifier classifies an input pattern vector and outputs the discriminated result with its score. An input pattern vector is obtained by analyzing 7 frames around the focusing frame and consists of 203 variables. For training the classifier, 7 frames around the manually-specified reference-point are used. Here we reduce the dimension of input vectors by statistically selecting 10～20 relevant variables before discrimination. Consonants are classified into 26 classes based on the canonical correlation analysis. In order to distinguish other phoneme-like parts from consonants, we add categories which represents vowels, noise, etc. In total, 34 classes listed in Table 2 are used.

Classification is performed by discriminant analysis. As the number of the classes increases, however, conventional discriminant analysis remarkably lowers its performance due to the use of common variables and a covariance matrix for all the classes. To conquer this defect, we have proposed pair-wise discrimination method[5] which discriminates multiple classes by combining the results of two-class discriminant analyses performed on the pairs of the classes. Here we adopt minimax method which classifies a pattern into the class whose minimum of a posteriori probabilities calculated on the pairs containing that class is maximum.

As the scores which the classfier gives the discriminated results, we used probability density explained in Section 2. Probability density is calculated by using common variables for all the classes, which does not necessarily matches the ones for pair-wise discrimination, intended to be the standard or absolute measure for multiple symbols at different frames.

Table 2 : Classes to be discriminated by phoneme classifier

consonants	?, p, py, t, k, ky, b, by, d, g, gy, m, n, ny, h, hy, s, sy, z, zy, ts, ch, r, w, y
vowels	a, i, u, e, o
others	mm(nasal-murmur), bz(buzz-bar), ns(noise)

? represents the forefront part of vowels preceded by no phonemes.
*y means palatalized consonants.

5. 2 HMM which treats phoneme-like symbol sequences

Using the phoneme-like symbol sequence as an input of HMM gives definite

meanings to the states of HMM, for example, the state of noise or the state of following vowel. The structure of HMM and the number of states are, therefore, decided top-down as follows :

- The initial state is set to represent the preceding vowel. In this paper, the preceding vowels are always noise since speech samples used are /CV/ syllables.
- The final state is set to represent the following vowel.
- A state of the consonant part around the reference-point is set.
- For the consonants preceded by buzz-bar or nasal-murmur, a state representing it is set just before the consonant part state.
- A transitive part state is set between the consonant part and the final state for some consonants.

As for state transition, we consider left-to-right models which allow only self-loops and single jumps. An example of HMM for *b* is illustrated in Fig. 3.

Each HMM is trained by the forward-backward algorithm[4]. Since we can presume the prevailing symbol in each state, for example, *bz* in buzz-bar state, adequate initial values of symbol output probabilities can be set and training is performed effectively.

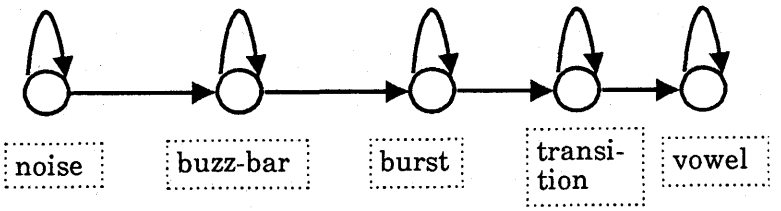

Fig. 3 A model for consonant /*b*/

6. EXPERIMENTAL RESULTS

Samples examined are all the Japanese consonants followed by one of the five Japanese vowels. The number of these /CV/ syllables is 101. Each syllable was uttered by 17～84 male speakers just once. Speech was recorded at the simple sound-proof booth and digitized to 12 bits at 18.5 kHz sampling rate. Every 10-ms frame of speech is analyzed to produce 28 variables representing spectrum envelope plus the mean square prediction error by 26-th order LPC analysis.

For training the phoneme classifier, namely discriminant analysis, all the 4013 acquired samples are used. For training HMM, 1222 samples (DS1) out of them are used, which means that about 50 samples are used to construct each consonant model. Another 202 samples (DS2), 2 for every /CV/ syllable, are used for testing.

Symbol sequences are segmented from several consecutive symbols of noise *ns*

to those of the following vowels.

6. 1 Recognition with scores by discriminant analysis alone

At first we made a recognition experiment using scores by discriminant analysis alone. The scores in the sequence are summed up for each consonant, and a pattern is classified into the one which got the highest summed score. The recognition rate was 86.6 % for DS1 and 87.6 % for DS2.

Recognition errors are examined in the Table 3. It is noticeable that there exists much confusion between consonants whose articulation places and articulation manners are different, for example, $g \rightarrow w$, $r \rightarrow h$. Not a few part of the confusion occurred, not because the classifier failed to discriminate around the reference-point, but because the classifier gave higher scores to incorrect consonant symbols distant from the reference-point.

Table 3 : content of recognition errors by discriminant analysis alone (DS1)

categories of errors	ratio
between consonants of the same articulation manner and place	12.3
between consonants of the same articulation manner	31.2
between consonants of the same articulation place	15.6
between consonants whose articulation manners and places are different	40.9
total of confusion between voiced consonants and those between unvoiced	65.6

6. 2 Recognition with scores by HMM alone

Next we made a recognition experiment using scores by HMM alone. This is the same as conventional VQ-based HMM except that HMM here treats phoneme-like symbols instead of VQ codes. The recognition rate was 88.6 % for DS1 and 76.2 % for DS2.

The contents of recognition errors are listed in the Table 4. Compared with Table 3, there occurred much more confusion between consonants of similar acoustic features, for example, $t \rightarrow p$, $d \rightarrow by$. This fact shows that HMM does not have enough power of accurate discrimination although it can classify patterns roughly.

Table 4 : content of recognition errors by HMM alone (DS1)

categories of errors	ratio
between consonants of the same articulation manner and place	10.1
between consonants of the same articulation manner	46.8
between consonants of the same articulation place	13.7
between consonants whose articulation manners and places are different	29.5
total of confusion between voiced consonants and those between unvoiced	74.1

6. 3 Recognition by combining two scores

Lastly we made a recognition experiment by combining the scores of discriminant analysis and HMM. The recognition rate is listed in Table 5 together with those of previous experiments. The confusion matrix is shown in Table 6.

Table 5 : recognition rate by each method (percent correct)

	DS1	DS2
discriminant analysis alone	86.8	87.6
HMM alone	88.6	76.2
integrated method	92.8	92.1
correct at the reference-point	88.1	88.6

The integrated method achieved the recognition rate of 92.8 % for DS1 and 92.1 % for DS2, which is higher by 5～15 % than the method using either discriminant analysis or HMM alone. The errors caused by the patterns distant from the reference-point, that occurred in using only discriminant analysis, and the confusions between similar consonants, that occurred in using HMM scores alone, were definitely reduced by combining two kinds of scores. These experimental results prove that discriminant analysis and HMM can compensate each other and the integrated method is effective.

7. CONCLUSIONS

A new recognition method is proposed. It integrates two stochastic methods : HMM, which grasps the global dynamic features, and discriminant analysis which discriminates based on the local detailed features. It conquered the defect of the two methods : HMM, which lacks the accuracy due to quantization of input vectors, and discriminant analysis, which assumes the precise detection of the reference-point.

The recognition experiment of all the consonants showed that this integrated method achieved higher recognition rates by 5～15 % than the case using either method alone.

In this paper, the experiment was performed for consonants in mono-syllables only. We are planing to study the application of our method to the consonants in the continuous speech, further considering the necessary symbols and the structure of HMM.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1972.

[2] L. A. Liporace. Maximum likelihood estimation for multivariate observations of Markov process. *IEEE Trans. Inf. Theory,* IT-28(5) : 729-734, 1982.

[3] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information

estimation of hidden Markov parameters for speech recognition. In *Proc. ICASSP*, pages 49-52, 1986.

[4] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Syst. Tech. J.*, 62(4) : 1035-1074, 1983.

[5] T. Kawahara, Y. Mizutani, S. Kitazawa, and S. Doshita. Application of pair-wise discrimination method to Japanese consonant recognition. *STUDIA PHONOLOGICA*, 22 : 83-93, 1988.